Exploring the UK Cyber Skills Gap through a mapping of active job listings to the Cyber Security Body of Knowledge (CyBOK)

Sam Attwood
SAttwood3@uclan.ac.uk
University of Central Lancashire
Preston, United Kingdom

Ashley Williams
ashley.williams@mmu.ac.uk
Manchester Metropolitan University
Manchester, UK

ABSTRACT

Background: The UK cyber skills gap/shortage amplifies the broader impact of cyber-attacks, which inflict harms such as privacy and economic loss on wider society. The demand is greatest (and growing fastest) in cyber-enabled disciplines, such as software engineering.

Objectives: In this paper, we create a term frequency-inverse document frequency representation of the Cyber Security Body of Knowledge (CyBOK). We then evaluate the potential of this representation by using it to automatically map job descriptions to the different areas of the CyBOK.

Method: We generate two representations of the CyBOK. The representations are mapped to a corpus of 454 job descriptions using TF-IDF. Comparing the similarity scores across these mappings allows us to identify relevant knowledge areas/groups.

Results: The results are preliminary, but suggest that the approach warrants further investigation. Certain job descriptions are mapped to certain knowledge areas/groups in a way that makes intuitive sense to the authors. However, there is a degree homogeneity to the scores returned for certain knowledge areas/groups. There are several threats to validity, most notably the low number of job descriptions that have been studied.

Conclusions: Our work shows that it is possible to automatically map job descriptions to the CyBOK in a meaningful way. Further research is required to address threats and to explore alternative mapping approaches. The authors intend to undertake this research culminating with a Grey Literature Informed Model of Practice in Secure Software Engineering.

CCS CONCEPTS

• **General and reference** → **Empirical studies**; • **Security and privacy** → **Software security engineering**; • **Software and its engineering** → **Designing software**; • **Social and professional topics** → **Computing education**.

KEYWORDS

Cyber Security, CyBOK, Secure Software Development, Higher Education, Grey Literature

EASE '23, June 14–16, 2023, Oulu, Finland
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0044-6/23/06.
https://doi.org/10.1145/3593434.3593459

ACM Reference Format:
Sam Attwood and Ashley Williams. 2023. Exploring the UK Cyber Skills Gap through a mapping of active job listings to the Cyber Security Body of Knowledge (CyBOK). In *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering (EASE '23), June 14–16, 2023, Oulu, Finland.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3593434.3593459

1 INTRODUCTION

Skills gaps between Higher Education (HE) and industry are prevalent and problematic (e.g., [4, 9, 12]). These gaps stifle economic output, growth, and profitability [1], and in the case of cyber security skills gaps, compound and amplify the broader and ever-present problem of cyber-attacks. Cyber-attacks and breaches can result in further economic loss; the 2021 UK Cyber Security Breaches Survey reporting that 39% of UK businesses disclosed a breach or attack over a 12-month period with the average cost of £8,460 [6]. In addition to this economic loss, cyber-attacks and breaches can inflict many other harms upon wider society, including privacy loss, dignity loss, and social detriment.

The skills gap between HE and industry with respect to Software Engineering (SE) is especially noticeable, with industry moving and innovating at a rapid pace, training itself through industrial qualifications and hyper-focused boot-camps, and favouring experience over academic achievement when hiring. This is compounded when the practice of secure software engineering is considered, with many academic curricula overlooking this aspect of SE. In the UK, this secure software engineering skills gap can be framed as a part of a broader cyber security skills gap where software engineers are cyber-enabled practitioners that often lack the skills needed to create and maintain secure software [21].

One of the key issues facing academia concerning SE and secure SE is understanding the current, and ever-evolving needs of industry so that it can develop relevant curricula. Industry panels and surveys can help to address this issue but could be complimented by an automated approach to better account for the fast changing needs of industry. In this article, we present our preliminary work towards automatically mapping UK technology-based job descriptions to the Cyber Security Body of Knowledge Project (CyBOK) [17]. Our aim is to facilitate immediate insights into the real-time needs of industry, and in doing so, allow: course designers to develop curricula which meets industry needs; local government and policy makers to measure the current skills gap; and funding bodies to identify areas of importance.

The paper investigates the following over-arching research question:

RQ1: To what degree can we measure the current cyber-skills gap in the UK through analysing live job adverts?

1.1 Contributions

The paper makes the following contributions:

- The CyBOK Mapping Reference v1.3 is made available in a structured JSON format.
- A term frequency-inverse document frequency (TF-IDF) based mapping approach for determining the CyBOK knowledge areas and groups most relevant to a given job description (or other document) is presented.
- We demonstrate the potential of this TF-IDF approach by interrogating an exemplar job market. Our emerging results show that the approach warrants further investigation and could ultimately be used to help: HE providers through more immediate insights into the needs of industry at any time; local governments and policy makers measure their cyber skills gaps; and funding bodies identify areas of importance.

2 RELATED WORK

The UK cyber security skills gap and shortage has been reported on extensively by the UK government [11, 15, 16, 21]. In the three most recent of these government reports, an analysis of Grey Literature (GL) in the form of job vacancies is presented [11, 15, 21]. In the most recent report, the locations, salaries, years experience, and skills associated with vacancies are all examined [21]. In contrast to this work, required skills are not mapped to the CyBOK as a part of the analysis, however it was used to identify search terms [21].

The CyBOK aims to systematise knowledge that is generally recognised as being related to cyber security [17]. The initial version was arrived at mainly through a series of consultation workshops (though online surveys, calls for position papers, and other techniques were also used) [17]. One of the key aims of the CyBOK is to support educators in designing cyber security curricula. At-least one prior study characterises curricula frameworks by mapping them to CyBOK knowledge areas [5] and it is now used as a tool to help certify degrees in the UK [14].

The CyBOK is complimented by other bodies of knowledge such as the IEEE Software Engineering Body of Knowledge (SWEBOK) [2]. The SWEBOK overlaps slightly with the CyBOK by including a section on 'Secure Software Development and Maintenance' [2]. The content in this section is akin to the content in the 'Secure Software Lifecycle' area of the CyBOK but less detailed. Additionally, the CyBOK has two additional knowledge areas related to software and platform security, which suggests it is the more mature body of knowledge in this regard.

Whilst the CyBOK covers software and platform security to a greater extent than the SWEBOK, the role of Software Engineer is best considered a cyber-enabled role. In the recent government report on cyber skills in the UK, cyber-enabled roles are said to be roles not widely recognised as cyber security roles, but roles that nonetheless require cyber security skills [21]. Software Engineer fits this definition. However, security remains overlooked in many SE curricula, with prior work having proposed modules and activities to help address this [20]. These are valuable but they fail to address the underlying problem of HE being unable to keep up with the changing state of secure SE practice in particular.

In this work, motivated by the UK cyber skills gap/shortage and security being overlooked in many SE curricula, we present an emerging approach for automatically mapping job descriptions to the CyBOK. In doing so we aim build on an existing body of work concerning the use of GL (such as job descriptions) as data and evidence in SE research [7, 19].

Galster *et al.* [3] surveyed New Zealand based job adverts for soft skills required by industry. Other than this study, we are aware of no research which has sought to quantify the skills-gap between academia and industry in SE.

3 METHODOLOGY

The methodology we used to map job descriptions to CyBOK knowledge areas is underpinned by a reference document [13]. This reference document is made available in a PDF format and lists terms (e.g., Abuse Cases) alongside their associated knowledge areas (e.g., Secure Software Lifecycle, denoted by the code SSL). The reference document notes that the knowledge areas listed for a given term are the knowledge areas most likely to contain said term; there is no guarantee that a term in the reference document is actually referenced in the CyBOK itself. A separate mapping framework document [13] explains how this reference document can be used alongside other resources to manually map UK degree programmes to the CyBOK.

As a first step towards an automated mapping method, we transformed the PDF reference document into an easier to work with JSON format. This was initially achieved using regex and a Python package, PyMuPDF[1]. The end result is a JSON reference in which a term and its associated knowledge areas exist as name/value pairs (e.g., `"ABUSE CASES": ["SSL"]`).

For the purposes of our analysis, we then transformed the JSON reference into two corpora:

(1) The first corpus consists of 21 documents. Each document in the corpus corresponds to a knowledge area. For each knowledge area, a document was created by concatenating all of the terms that have the said knowledge area associated with it. This means that terms in the original reference can contribute to multiple documents in the corpus (e.g., a name/value pair of `"ACCESS TOKEN": ["WAM", "HF", "F", "AAA"]` will result in `"ACCESS TOKEN"` being added to multiple documents corresponding to the Web Application & Mobile Security, Human Factors, Forensics, and Authentication, Authorisation, & Accountability knowledge areas).

(2) The second corpus consists of 5 documents. Each document in the corpus corresponds to a group of related knowledge areas. These groups were derived from the main CyBOK page and the knowledge areas that contribute to each group are shown in table 1. Note that the 'Introductory concepts' group in the CyBOK, which contains just an 'Introduction to CyBOK' knowledge area, does not have a corresponding document in the corpus. For each group, a document was created by concatenating all of the terms that have a contributing knowledge area associated with it. If a term had multiple

[1] https://pymupdf.readthedocs.io/en/latest/index.html

knowledge areas belonging to the same group associated with it, the term was added to the document multiple times. Equally, the same term can contribute to multiple groups if it has associated knowledge areas that belong to different groups.

After removing English stopwords from both of the corpora using the Natural Language Toolkit [8], we then calculated the term frequency-inverse document frequency (TF-IDF) representation for each as follows:

$$w_{i,j} = tf_{i,j} * log(N/df_i)$$

Where $w_{i,j}$ is the calculated TF-IDF for the i^{th} token in the j^{th} document, $tf_{i,j}$ is the frequency of the i^{th} token in the j^{th} document, N is the total number of documents, and df_i is the number of documents that contain the i_{th} token. This was achieved using the Gensim Python package [18], which was also used to run similarity queries against both of the TF-IDF representations. The representations are referred to as *ka_tf-idf* and *kg_tf-idf* for the remainder of this article, with the former (*ka_tf-idf*) being a representation of the first 21 document corpus and the latter (*kg_tf-idf*) being a representation of the second 5 document corpus.

The job descriptions we used to query the *ka_tf-idf* and *kg_tf-idf* representations (see section 5) were retrieved from a freely accessible API[2]. All of the descriptions available as of 21-02-2023 were retrieved. In addition to the job descriptions, we also retrieved the following using this API: title, minimum salary, maximum salary, and location. To facilitate meaningful analysis, we grouped the job records by labelling them with a simplified title (job type). The first and second author both labelled the job records independently and came together to resolve disagreements and consolidate.[3]

4 EXPLORATORY ANALYSIS OF THE JOB DESCRIPTIONS

In this section we report exploratory analysis of retreived job records to help highlight limitations in our work so far. 454 job records were retrieved from the API, with each retrieved record having a title, description, minimum salary, maximum salary, and location fields. These were then categorised into job types to facilitate meaningful analyses. Overall, this analysis supports the use of this API at this initial stage and demonstrates the records are broadly representative of cyber and cyber-enabled jobs in the UK. However, the use of a single API and two annotators remains a limitation of our work at this stage as discussed in Section 6.1.

Figure 1 shows a heat-map of the locations associated with job records. Regions with many jobs are highlighted by the denser areas on the heat-map shown in red and orange. Regions with fewer jobs are shown in yellow and locations with no jobs show no colour. From figure 1 it is therefore apparent that the majority of the retrieved records represent jobs in and around the Greater London region, with 191 records being in the Greater London region itself. This is consistent with the analysis presented in most recent cyber skills report from the Government [21]. Greater Manchester is the second most represented region with 32 records.

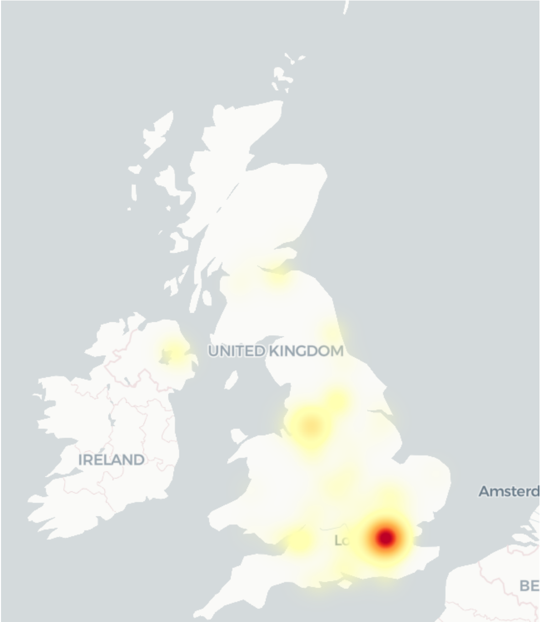

Figure 1: A heat-map showing job locations.

Table 2 shows the categorised job types, and the number of records that have been labelled with each of these titles. It shows that eight categories were decided upon by the authors: 'Dev', 'Other', 'Manager', 'DevOps', 'Security', 'Data', 'QA', and 'IT'. Most of the records have been assigned to the 'Dev' category, which includes roles such as 'iOS Developer (Remote, UK Based)', 'Backend Engineer', and 'Front-End Developer'. The greater quantity of 'Dev' is most likely a symptom of using a single API for a job board that places an emphasis on software developer roles[4]. However, this could also be a symptom of the annotators/authors backgrounds, as both have backgrounds in SE and are therefore able to recognise these types of roles with greater ease (c.f. recognising an 'IT' role as distinct from a 'DevOps' or 'Other' type of role).

Figure 2 shows the mean salary associated with the different job types. Each point shown in figure 2 represents a different record with the mean salary being calculated using the minimum and maximum salary values given in the description. Jobs inside and outside Greater London have been separated using the y-axis and while there is an expected difference in mean salary inside and outside Greater London (note our plot does not account for differences in cost of living), there is comparatively little difference in mean salary across job types.

5 MAPPING JOB DESCRIPTIONS TO KNOWLEDGE AREAS

Job records were automatically mapped to CyBOK knowledge groups and areas by using their descriptions to query *kg_tf-idf* and *ka_tf-idf.* Figure 3 shows the distribution of similarity scores returned across different knowledge groups when querying *kg_tf-idf.* The eight job types are shown separately (in different colours) to

[2] https://devitjobs.uk/job_feed.xml

[3] Available: https://github.com/samattwood9/emerging-jobs-cybok.

[4] https://devitjobs.uk/about

Table 1: Groups, contributing knowledge areas, and their short names.

Group	Knowledge areas	Short name
Human, org., & reg.	Risk management & governance	RMG
	Law & regulation	LR
	Human factors	HF
	Privacy and online rights	POR
Attacks & defences	Malware & attack technologies	MAT
	Adversarial behaviours	AB
	Security operations & incident management	SOIM
	Forensics	F
Systems security	Cryptography	C
	Operating systems & virtualisation	OSV
	Distributed systems security	DSS
	Formal methods for security	FM
	Authentication, authorisation, & accountability.	AAA
Software & platform	Software security	SS
	Web & mobile security	WAM
	Secure software lifecycle	SSL
Infrastructure security	Applied cryptography	AC
	Network security	NS
	Hardware security	HS
	Cyber physical systems	CPS
	Physical layer & telecommunications	PLT

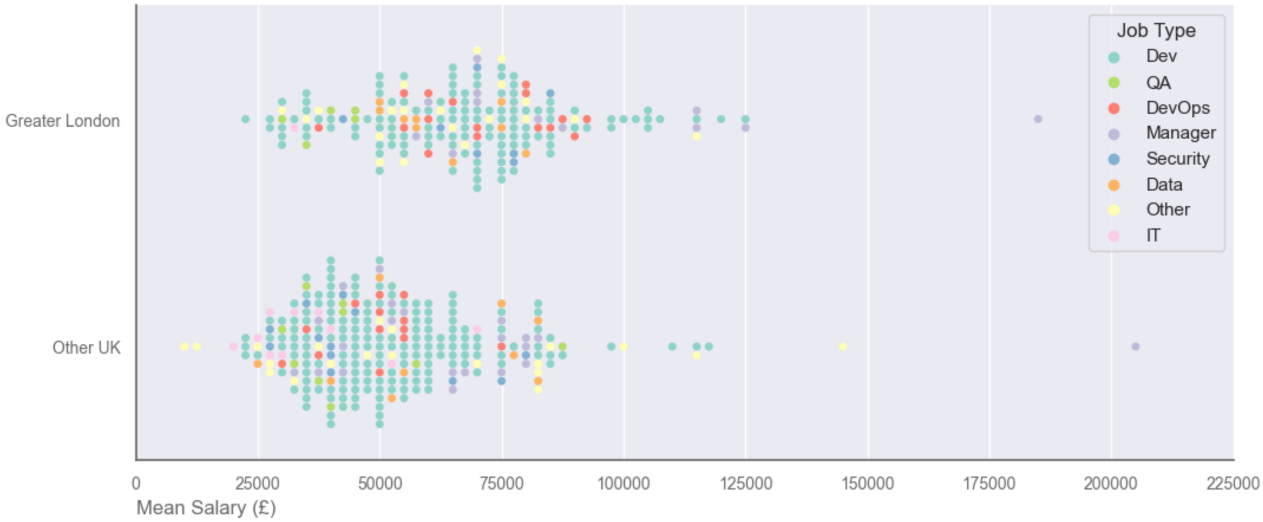

Figure 2: A categorical scatterplot showing mean salary (£) for different job types inside and outside Greater London.

allow for the similarity scores and the legitimacy of our approach to be sense-checked at this early stage (e.g., if the approach has potential we would expect to see 'Dev' type jobs score higher in the 'Software & platform' group than 'IT' type jobs, similarly we would expect to see 'Security' type jobs score higher in the 'Attacks & defences' group than 'Dev' type jobs).

Overall, Figure 3 suggests our approach has potential and warrants further development; certain job types score higher with respect to certain groups in a way that makes intuitive sense (e.g., 'Dev' roles do score higher in the 'Software & Platform' group than 'IT' type jobs, and 'Security' type jobs score higher in the 'Attacks & Defences' group than 'Dev' type jobs). However, Figure 3 also highlights a degree of homogeneity in the similarity scores that are returned when querying *kg_tf-idf.* This is especially apparent when examining the 'Infrastructure' and 'Systems' groups, in which the similarity scores vary only small amounts across different job types.

Analysis performed after using the same descriptions to query *ka_tf-idf* produced similar findings. There was a degree of homogeneity to similarity scores returned for certain knowledge areas (e.g., 'Malware & attack technologies' and 'Cryptography'). However, as with the knowledge groups, certain job types did score higher with certain knowledge areas in a way that makes intuitive

Table 2: Number of records by job type category.

Job Type	Number of records
Dev	289
Other	40
Manager	33
DevOps	29
Security	19
Data	18
QA	14
IT	11

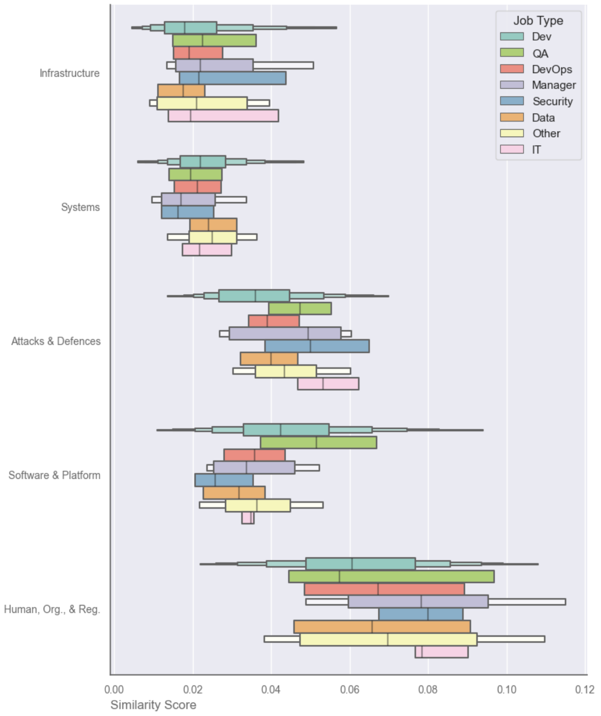

Figure 3: A boxenplot showing the similarity score for job types across the main 5 knowledge groups in the CyBOK.

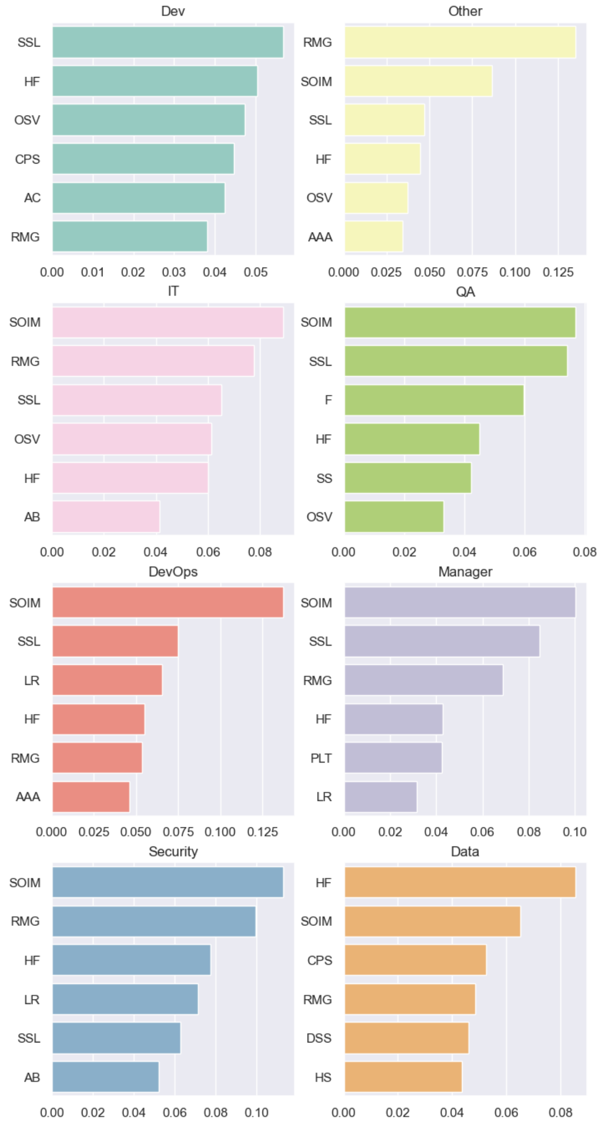

Figure 4: Bar charts showing the similarity scores of the top six scoring knowledge areas for each illustrative record.

sense. For example, 'Dev' type jobs scored greater than other types in the 'Web & mobile security' knowledge area, and lesser than other types in the 'Security operations & incident management' knowledge area.

To further explore the mappings we selected illustrative job records and explore their mappings in greater detail. One record was selected for each job type: Dev ⟶ Full Stack Developer - Fully Remote, Other ⟶ Data Governance Specialist, Manager ⟶ Agile Delivery Manager, DevOps ⟶ Clouds DevOps Engineer, Security ⟶ IT Security Analyst, Data ⟶ Senior Data Scientist, QA ⟶ Software Test Engineer, IT ⟶ IT Support. The illustrative records were selected based on their unique and descriptive job titles.

Figure 4 shows the similarity score for each illustrative job description across the six highest scoring knowledge areas. These mappings further evidence the potential of the approach with the highest scoring knowledge areas for the illustrative job descriptions making intuitive sense (e.g., the 'Dev' description scores highest in the 'Secure software lifecycle' knowledge area and the 'Other' description, which corresponds to a 'Data Governance Specialist' role, scores highest in the 'Risk management & governance' knowledge areas). The presence of the 'Applied cryptography' knowledge area in the six highest scoring areas for the 'Dev' role is notable and at first does not make intuitive sense given the absence of the 'Web & mobile security' knowledge area. However, further analysis of the illustrative description shows that the role will be supporting the development of a cryptocurrency platform, which arguably justifies the inclusion of the 'Applied cryptography' area.

6 DISCUSSION AND CONCLUSIONS

Section 3 clearly demonstrates that it was possible to create a TF-IDF based approach to mapping natural language data to the CyBOK. Establishing how to transform the PDF version of the Mapping Reference into a structured JSON format was the main challenge that needed to be overcome to achieve this. Now that this has been achieved, researchers should be able to explore alternative approaches with greater ease.

In terms of whether TF-IDF based approaches (specifically) warrants further investigation, we argue that they do. From section 5 it is clear that the two complimentary approaches described in this article produce mappings that make intuitive sense. This is true when job descriptions of a given type are considered collectively and also when considered in isolation. However, it is also true that there is a degree of homogeneity to the similarity scores that are produced (e.g., the 'Human Factors' knowledge area and the 'Human, Org., & Reg.' group score highly for many of the job descriptions with little difference between descriptions of different job types), which may limit the potential of the approach as a means of classifying job descriptions and other GL in the future.

6.1 Threats to validity

The results presented in this paper are preliminary and on-going. As such, there are a number of threats that need to be addressed: a relatively low number of job descriptions 454 have been considered. Furthermore, all of these job descriptions were retrieved using the same API. The validity of our findings could be improved by collecting a greater quantity of job records from a variety of APIs; two annotators with similar backgrounds (in SE) annotated and categorised the job descriptions. The validity of our findings could be improved by adopting a more systematic annotation process and using a greater number of annotators; the TF-IDF based mapping approaches have not been compared to any alternatives that may provide a superior performance. Moreover, there is not currently a systematic way of comparing performance. To do this and enhance the validity of our findings, a new data-set with GL annotated specifically for CyBOK knowledge areas and groups could be created.

6.2 Future research

The threats highlighted indicate several issues that will be addressed through future research. In a follow-up study we intended to collect a greater quantity of GL from a variety of sources (e.g., practitioner blog articles, and descriptions of academic curricula). This follow-up study will also make use of multiple annotators to create a data-set that is annotated specifically for CyBOK knowledge areas and groups. The authors intend to achieve this in close collaboration with cyber and cyber-enabled practitioners. Finally, this better quality data-set will enable multiple mapping approaches to be compared.

In addition, this paper serves as preliminary work towards the development of a GL Informed Model of Practice in Secure Software Engineering (GLIMPSE). This model will be descriptive and comparable to the Building Security in Maturity Model (BSIMM) [10]. A GLIMPSE is motivated by the observation that GL could facilitate the creation of a more complete descriptive model that considers a greater quantity and variety of perspectives. Like the CyBOK and SWEBOK a GLIMPSE would benefit educators by helping them design curricula. Furthermore, a GLIMPSE would benefit practitioners by highlighting the most relevant practices to them, and therefore contribute towards greater overall software security.

REFERENCES

[1] 2022. *Business Barometer 2022. Navigating the skills landscape.* Technical Report. The Open Univerisity, British Chambers of Commerce.

[2] P Borque and R Fairley. 2014. *Guide to the Software Engineering Body of Knowledge, Version 3.0, IEEE Computer Society.* Technical Report.

[3] Matthias Galster, Antonija Mitrovic, Sanna Malinen, and Jay Holland. 2022. What Soft Skills Does the Software Industry Really Want? An Exploratory Study of Software Positions in New Zealand. In *Proceedings of the 16th ACM IEEE International Symposium on Empirical Software Engineering and Measurement.* 272–282.

[4] Christopher J Hahn and Jeanine E Gangeness. 2019. Business, Leadership And Education: A Case For More Business Engagement In Higher Education. *American Journal of Business Education* 12, 1 (2019), 19–31.

[5] Joseph Hallett, Robert Larson, and Awais Rashid. 2018. Mirror, mirror, on the wall: What are we teaching them all? Characterising the focus of cybersecurity curricular frameworks. In *USENIX Workshop on Advances in Security Education.*

[6] Emma Johns. 2021. *Cyber Security Breaches Survey 2021.* Technical Report. Department for Digital, Culture, Media, and Sport.

[7] Barbara Kitchenham, Lech Madeyski, and David Budgen. 2022. How Should Software Engineering Secondary Studies Include Grey Material? *Transactions on Software Engineering* (2022).

[8] Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028* (2002).

[9] Jaser K Mahasneh and Walid Thabet. 2015. Rethinking construction curriculum: A descriptive cause analysis for the soft skills gap among construction graduates. In *51st ASC Annual International Conference Proceedings.* 35.

[10] Gary McGraw. 2015. Software Security and the Building Security in Maturity Model (BSIMM). *J. Comput. Sci. Coll.* 30, 3 (jan 2015), 7–8.

[11] Darragh McHenry, Tania Borges, Alex Bollen, Jayesh Navin Shah, Sam Donaldson, David Crozier, and Steven Furnell. 2021. *Cyber security skills in the UK labour market 2021.* Technical Report. Department for Digital, Culture, Media, and Sport.

[12] Steven Miller. 2014. Collaborative approaches needed to close the big data skills gap. *Journal of Organization design* 3, 1 (2014), 26–30.

[13] Lata Nautiyal, Joseph Hallett, James Clements, Benjamin Shreeve, and Awais Rashid. 2021. *CyBOK Mapping Reference Issue 1.3.0.* Technical Report.

[14] Lata Nautiyal, Awais Rashid, Joseph Hallett, and Ben Shreeve. 2020. *The UK's Cyber Security Degree Certification Programme: A CyBOK Case Study.* Technical Report.

[15] Daniel Pedley, Tania Borges, Alex Bollen, Jayesh Navin Shah, Sam Donaldon, Steven Furnell, and David Crozier. 2020. *Cyber security skills in the UK labour market 2020.* Technical Report. Department for Digital, Culture, Media, and Sport.

[16] Daniel Pedley, Darragh McHenry, Helen Motha, and Jayesh Navin Shah. 2018. *Understanding the UK Cyber Security Skills Labour Market.* Technical Report. Department for Digital, Culture, Media, and Sport.

[17] Awais Rashid, George Danezis, Howard Chivers, Emil Lupu, Andrew Martin, Makayla Lewis, and Claudia Peersman. 2018. Scoping the Cyber Security Body of Knowledge. *IEEE Security & Privacy* 16, 3 (2018), 96–102. https://doi.org/10.1109/MSP.2018.2701150

[18] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.* ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[19] Ashley Williams and Jim Buchan. 2022. Using the case survey methodology for finding high-quality grey literature in software engineering. In *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering 2022.* 1–9.

[20] Xiaohong Yuan, Li Yang, Bilan Jones, Huiming Yu, and Bei-Tseng Chu. 2016. Secure software engineering education: Knowledge area, curriculum and resources. *Journal of Cybersecurity Education, Research and Practice* 2016, 1 (2016), 3.

[21] Gabriele Zatterin, Grace Atkins, and Jayesh Navin Shah. 2022. *Cyber security skills in the UK labour market 2022, Technical Report.* Technical Report. Department for Digital, Culture, Media, and Sport.